Graham_HierarchicalClusteringProject

Roland Graham

2023-09-01

Using Agglomerative Hierarchical Clustering to Identify Mall Customer Segments

The following is a fictitious dataset regarding the customer demographics of a hypothetical mall. The dataset contains 200 observations, each from an imaginary customer. The variables are demographical information on the consumers which are

1. Age (in years)
2. Income (in thousands of dollars)
3. Spending score (measures the spending nature of the person on a scale of 1-100 where 1 is spending the least money and 100 is spending the most money)

Source: https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python

Our goal is to divide the 200 consumers into distiinct demographical segments based on the three characteristics mentioned above. For example, we could possibly have a segment that is primarily of younger people, high incomes, and high spending scores or a segment of older people, low incomes, and low spending scores.

It is important to the mall that we identify the segments that have the highest spending scores as that will be the mall's target audience. We could then relay that information to the product management and marketing departments and tell them to cater to those select demographics of people.

We can accomplish this through clustering, which is where we divide the dataset into groups based on characteristics rather than using them to arrive at an outcome value.

Hierarchical clustering is a form of clustering that builds clusters in a stepwise format. It is ideal for data that has a low number of observations and variables, which is the case for the data here. There are two main categories of hierarchical clustering:

1. Agglomerative: Each observation starts as its own cluster and similar observations are merged together in order to form a set number of final clusters in the end
2. Divisive: All observations start as one giant cluster and dissimilar observations break away until there are are a set number of final clusters in the end

We will be using the agglomerative method as its the most popular. It should also be noted that hierarchical clustering cannot support a mix of continuous and binary variables. This means that a variable such as "Gender" which can take a value of Male or Female, must be removed from the dataset since the rest of the variables are continuous metrics.

```r
# Loads the factoextra library
library(factoextra)
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

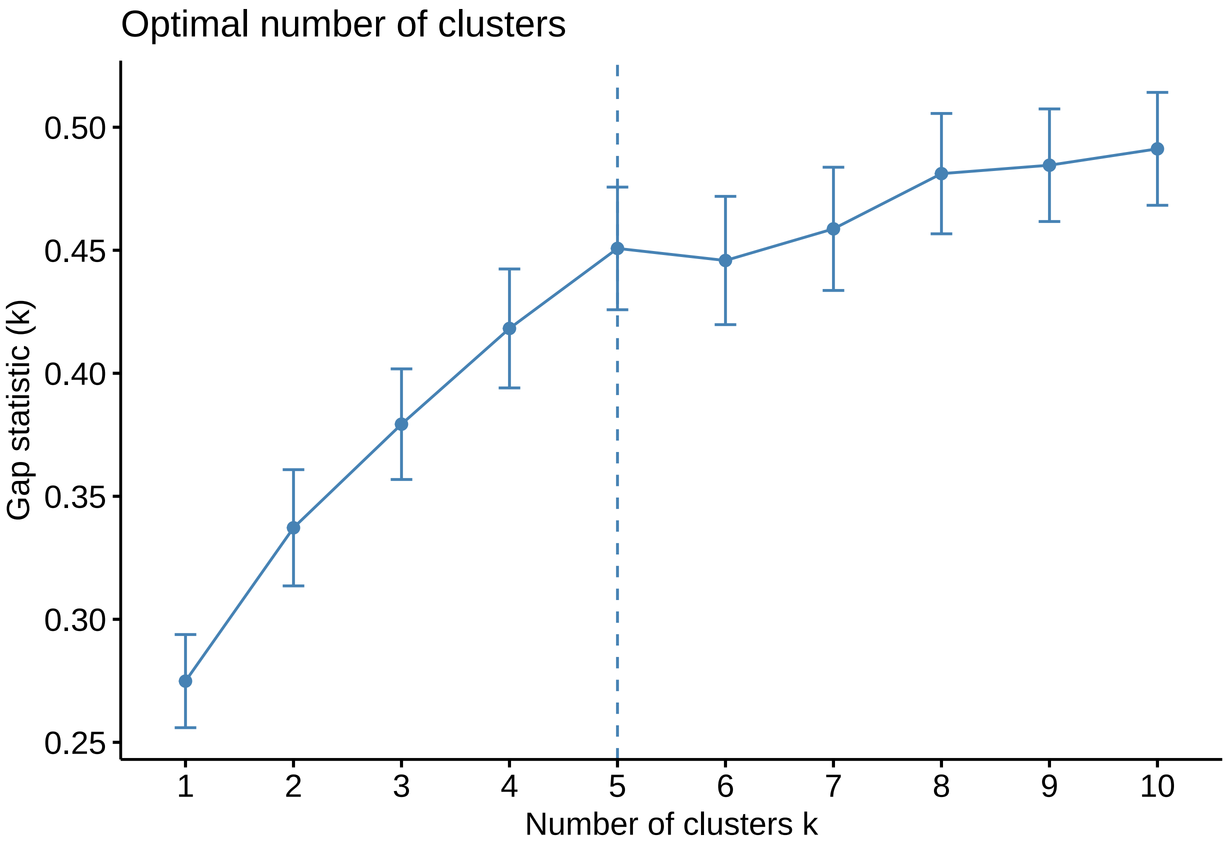
```
#Runs the data and adds simpler column names to the Income and Spending Score
# variables
data = read.csv("/Users/rolandgraham/Documents/Mall_Customers.csv", header = TRUE)
colnames(data)[4] = "Income"
colnames(data)[5] = "Score"

# Removes the CustomerID and Gender variables
# Note: We have to remove Gender because hierarchical clustering only works with
# continuous variable types
data = data[,3:5]

# Standardizes the variables
data$Age = (data$Age - min(data$Age)) / (max(data$Age) - min(data$Age))
data$Income = (data$Income - min(data$Income)) / (max(data$Income) - min(data$Income))
data$Score = (data$Score - min(data$Score)) / (max(data$Score) - min(data$Score))
```

The one thing we have to determine, however, is how many clusters will be in the final result. One way this can be accomplished is with the Gap Statistic Method, which creates sets of variables with different numbers of clusters and calculates the average distance between the variables. We want a model with a high Gap Statistic as that means the clusters will be most distinct from one another, but we also want a model with the fewest number of clusters as simpler models are better than complex models. The key is to find an "elbow" in the Gap Statistic graph where there is a dropping off point in terms of the increase in the Gap Statistic as we keep adding clusters.

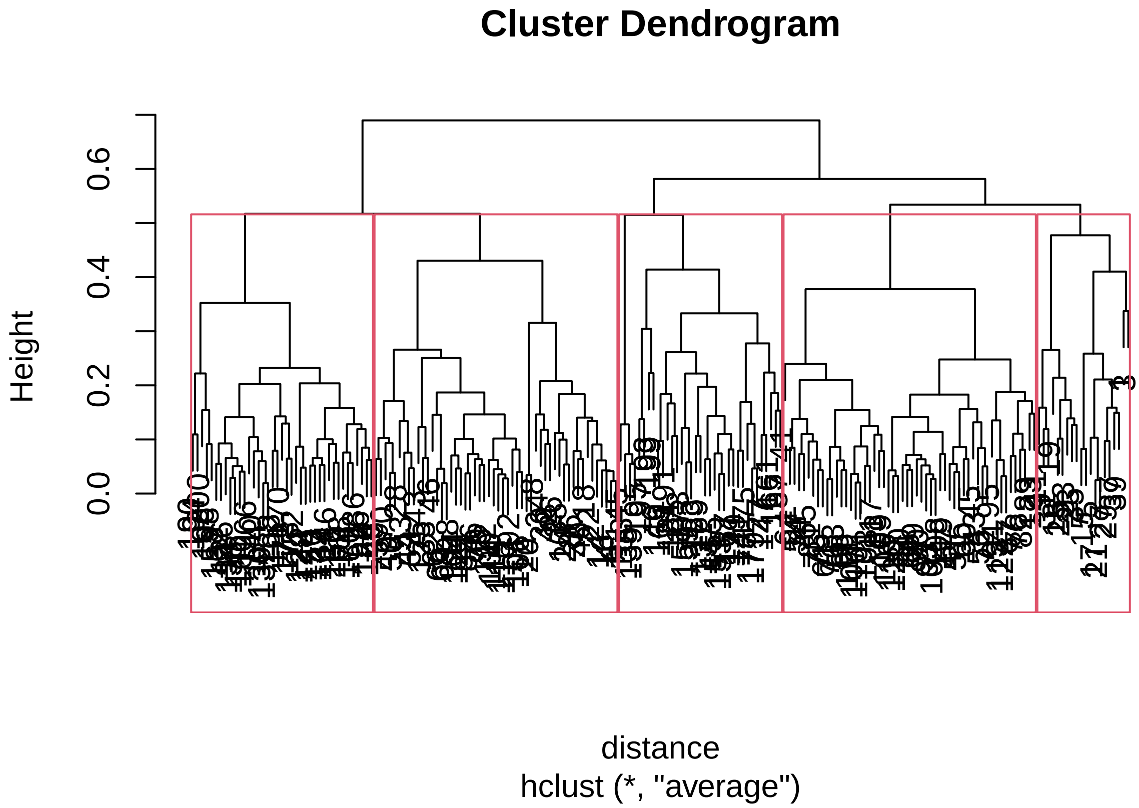
```
# Uses the gap stat method to determine the ideal number of clusters
# The plot shows the optimal number is k = 5 clusters
set.seed(35)
fviz_nbclust(data, hcut, method = "gap_stat")
```



From the graph, we can see that the optimal number of clusters is five. Now, we can perform agglomerative hierarchical clustering. The way we can map out the hierarchical clustering process is with a tree-like diagram called a dendrogram.

```
# Calculates the average distance for each hierarchical cluster
set.seed(35)
distance = dist(data, method = "euclidean")
hier.average = hclust(distance, method = "average")

# Creates and plots a dendrogram to show the two clusters
plot(hier.average)
rect.hclust(hier.average, k = 5)
abline(h = 5, col = "red")
```



```
# Identifies which variables are in which cluster
cluster.cut = cutree(hier.average, k = 5)

# Adds the cluster identifier to the dataset
data$Cluster = cluster.cut
```

The observations are at the bottom of the tree and the five clusters are separated by the red lines. We can then compute summary statistics for the five clusters.

```
# Un-standardizes the data in order to put the variables back into their original units
data = read.csv("/Users/rolandgraham/Documents/Mall_Customers.csv", header = TRUE)
colnames(data)[4] = "Income"
colnames(data)[5] = "Score"
data = data[,3:5]
data$Cluster = cluster.cut
```

```
# Shows the summary statistics for the first cluster
summary(data$Age[data$Cluster == 1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   19.00   35.00   43.50   43.90   53.25   67.00
```

```
summary(data$Income[data$Cluster == 1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   15.00   19.00   23.50   24.45   29.25   37.00
```

```
summary(data$Score[data$Cluster == 1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    3.00    6.00   14.50   19.10   31.25   40.00
```

```
# Shows the summary statistics for the second cluster
summary(data$Age[data$Cluster == 2])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   18.00   21.00   23.50   24.65   28.25   35.00
```

```
summary(data$Income[data$Cluster == 2])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   15.00   28.00   41.00   42.94   60.00   76.00
```

```
summary(data$Score[data$Cluster == 2])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   29.00   49.75   58.00   62.08   75.25   99.00
```

```
# Shows the summary statistics for the third cluster
summary(data$Age[data$Cluster == 3])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   32.00   47.25   50.00   53.26   63.00   70.00
```

```
summary(data$Income[data$Cluster == 3])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   38.00   47.25   54.00   54.20   62.00   69.00
```

```
summary(data$Score[data$Cluster == 3])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   28.00   43.00   48.00   48.56   55.00   60.00
```

```
# Shows the summary statistics for the fourth cluster
summary(data$Age[data$Cluster == 4])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   27.00   30.00   32.00   32.69   35.50   40.00
```

```
summary(data$Income[data$Cluster == 4])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   69.00   75.50   79.00   86.54   95.00  137.00
```

```
summary(data$Score[data$Cluster == 4])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   63.00   74.50   83.00   82.13   90.00   97.00
```

```r
# Shows the summary statistics for the fifth cluster
summary(data$Age[data$Cluster == 5])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   19.00   35.00   43.00   41.69   47.50   59.00
```

```r
summary(data$Income[data$Cluster == 5])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   71.00   77.50   85.00   88.23   97.50  137.00
```

```r
summary(data$Score[data$Cluster == 5])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00   10.00   16.00   17.29   23.50   39.00
```

From comparing the five clusters, we can see that the clusters with the two highest spending scores were

1. People in their 30s with high income
2. People in their 20s with low to moderate income

The cluster with the third highest spending score was

3. People in their 50s with moderate income

And the clusters with the two lowest spending scores were

4. People in their 40s with low income
5. People in their 40s with high income

This means that from an age perspective, younger people tend to spend at this mall more than the older people given the divide in their spending scores. In terms of income, the two clusters are people who make $86,540 per year and $42,940 respectively. This means that products can be marketed to people whose income falls in the broad range of $40000-$90000 per year.